# 修 士 論 文 の 和 文 要 旨

| | | |
|---|---|---|
| 大学院　電気通信学　研究科　　博士前期課程　　システム工学　　専攻 | | |
| 氏　　名 | 小熊　淳一 | 学籍番号 0535010 |
| 論　文　題　目 | 話題の共通度に基づく文書クラスタリング | |

要　　旨

今日，Web 等を利用して電子化された文書を容易に入手できるようになった．そのため必要とする文書を効率的に得るために情報検索技術が必要となっている．そこで，文書クラスタリング技術が盛んに検討されている．

文書クラスタリングを用いてWeb 検索を支援する場合，対象とする文書集合にはある共通の話題が存在すると考えられる．共通話題が存在する文書集合では，共通話題以外の類似性を用いてクラスタリングを行うべきであると考えられる．しかし既存の手法では共通話題の存在を仮定しておらず，共通話題がノイズとなってクラスタリング精度が低下してしまう可能性が考えられる．したがって，共通話題を除去するための新たな手法が必要となる．

文書集合の共通話題の内容が狭い分野に限定される場合には，広い分野にまたがる場合よりも，文書集合はより多くの共通性を持つと考えられる．したがって，除去するべき文書集合の共通話題は，文書集合の共通性の度合から知ることが出来ると考えられる．そのためには，階層的に分類された文書集合の共通性を数値化する技術が必要となる．

そこで本研究では，階層的に分類された文書集合の共通度合を求める手法を提案する．そして文書集合から共通度合を用いて抽出した共通話題を除去してクラスタリングを行う手法を提案し，既存手法との比較を行った．

既存の共通度合を数値化する手法と比較して，提案した文書集合の共通度合を計測する手法は階層構造を持つ文書集合の共通性を表現するのに妥当な手法であった．また，既存のクラスタリング手法であるtf・idf法，GMM+EM法よりも，類似文書のクラスタリング手法として有効であった．